心理科学进展 2021, Vol. 29, No. 12, 2272–2280
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2021.02272

认知诊断 Q 矩阵估计(修正)方法

李 佳 毛秀珍 张雪琴

(四川师范大学教育科学学院, 成都 610066)

摘 要 Q 矩阵代表着项目考察的属性, 反映了项目的重要特征, 其正确性是影响认知诊断分类准确性的关键因素。研究 Q 矩阵估计(修正)方法具有重要价值。首先, 研究从是否采用认知诊断模型将 Q 矩阵估计(修正)分为基于认知诊断模型视角下的参数化方法和基于统计视角下的非参数方法。然后, 分别从最优项目质量、最优模型数据拟合和参数估计视角对它们进行分类介绍, 评析不同方法的特征和表现、区别与联系、优势与不足。最后, 提出几个未来研究问题: 在复杂测验条件下系统比较各种方法; 校准知识状态和参数估计误差、结合多种思路和方法等多角度提出 Q 矩阵估计(修正)方法; 研究多级评分项目、混合测验模型、属性多级、属性个数未知甚至 Q 矩阵元素为连续变量等条件下的 Q 矩阵估计(修正)方法。

关键词 认知诊断模型, Q 矩阵, Q 矩阵估计(修正)方法, 数据拟合, 参数估计
分类号 B841

1 引言

2020 年 6 月 30 日, 中共中央全面深化改革委员会审议通过了《深化新时代教育评价改革总体方案》。它明确要求教育评价要“改进结果评价, 强化过程评价, 探索增值评价, 健全综合评价”。认知诊断结合了认知心理学和现代测量学知识, 通过考生的作答反应分析其潜在的认知技能和心理加工过程, 不仅能提供详细和全面的诊断信息, 还有利于因材施教和实现个性化学习。于是, 认知诊断评估(cognitive diagnosis assessment, CDA)在新时代教育评价改革背景下将得到越来越广泛的关注与深入研究。

Q 矩阵表征了项目与属性的关系, 是认知诊断研究与实践应用的基础。有研究表明: 标定有误的 Q 矩阵会增大参数估计误差并降低诊断分类正确率(Rupp & Templin, 2008; de la Torre, 2009; 涂冬波 等, 2012)。于是, Q 矩阵的正确性关系到认知诊断结果的准确性和可靠性, 如何获得准确的 Q 矩阵具有重要研究意义。实践中, 研究者常常根据学生口语报告和领域专家分析来标定测验 Q 矩阵或者依据作答反应数据来估计测验 Q 矩阵。前者主要运用质性分析方法, 具有一定主观性; 后者基于数据分析获得 Q 矩阵, 也可能不符合真实情况。喻晓锋、罗照盛、高椿雷等人(2015)提出先估计再修正, 通过对项目 q 向量的“双重修订”可以保证测验 Q 矩阵的标定效率。事实上, 无论在专家标定 Q 矩阵的基础上通过数据分析进行修正, 还是专家对估计的 Q 矩阵进一步修正, 都是可行的“双重修订”模式。由此可见, Q 矩阵估计或 Q 矩阵修正是标定 Q 矩阵中不可或缺的重要步骤。事实上, Q 矩阵估计与 Q 矩阵修正的方法是相通的, 二者的区别在于是否以预估的 Q 矩阵为前提, 即前者基于专家界定的部分项目的 q 向量和作答反应或仅基于作答反应估计测验 Q 矩阵, 而后者则对专家界定的初始 Q 矩阵进行检验或修正。因此, 本研究统称为 Q 矩阵估计(修正)方法。

近 10 年来, 针对 Q 矩阵估计(修正)问题, 研究者们开展了深入研究并提出了大量方法。把握这些方法的思想和步骤, 解析不同方法之间的区别和联系, 厘清它们的特点、优势与不足, 不仅能呈现 Q 矩阵估计(修正)方法的研究脉络和发展方向, 还能为实践者选用恰当方法提供依据。因此, 梳理 Q 矩阵估计(修正)方法具有重要价值和意义, 也成为本文的核心内容。

收稿日期: 2021-04-02
通信作者: 毛秀珍, E-mail: maomao_wanli@163.com

Q 矩阵是建立可观察的作答反应和不可观察的项目特征与被试知识状态(knowledge states, KS)之间关系的桥梁。根据是否采用认知诊断模型(cognitive diagnosis models, CDMs)表征它们的关系，研究将 Q 矩阵估计(修正)方法分为参数化和非参数方法两个大类，见表 1。第二、第三部分分别从最优项目区分度、模型数据拟合和参数估计角度进一步对它们进行梳理、分类、介绍和评析；第四部分进行总结和展望。

下文用 N、J 和 K 分别表示被试人数、项目个数和测验考察的属性个数。i 、 j 和 t 分别表示被试、项目和项目可能的得分值。$\boldsymbol{\alpha}$ ($\hat{\boldsymbol{\alpha}}$)和 $\boldsymbol{\beta}$ ($\hat{\boldsymbol{\beta}}$)分别表示真实(估计)的 KS 和项目参数， $\boldsymbol{q}_j$ 表示项目 j ($j=1,2,\cdots,J$)的属性模式， $\boldsymbol{q}_{jc}$ $(c=1,2,\cdots,2^K-1)$ 表示项目 j 的第 c 类候选属性模式。$\boldsymbol{Y}$ 和 $\tilde{\boldsymbol{Y}}$ 分别表示观察反应向量和基于模型的期望反应向量，$\boldsymbol{\eta}$ 表示没有失误和猜测条件下的理想反应向量。Y_{ij} 、 $\tilde{Y}_{ij}$ 和 η_{ij} 分别表示被试 i 在项目 j 上的观察、期望和理想反应。

2　参数化 Q 矩阵估计(修正)方法

Q 矩阵与项目特征和数据拟合有密切关系。一般地，正确的 Q 矩阵应使项目区分度最高、模型与数据拟合最好。于是，从所有有限多种可能的 q 向量中选择使项目区分度最高或模型数据拟合最好的属性模式作为项目 q 向量，是早期 Q 矩阵估计(修正)方法的出发点。此外，Q 矩阵元素还可以视作未知参数，通过参数估计方法估计而得。

2.1　最优项目区分度方法

从所有可能的属性模式中选择具有最优项目区分度的属性模式作为项目 q 向量是最优项目区分度方法的核心思想。这类方法包括 δ 法(de la Torre, 2008)、γ 法(涂冬波 等, 2012)、ς^2 法(de la Torre & Chiu, 2016)和 stepwise 法(Ma & de la Torre, 2020)。

2.1.1　项目鉴别力指数：δ 方法

de la Torre (2008)根据鉴别力指数的定义，提出选择使项目 j 中高低分组被试正确作答概率之差最大的属性模式作为它的 q 向量，称为 δ 方法。以 DINA 模型为例，对项目 j 而言，首先根据被试是否掌握候选 $\boldsymbol{q}_{jc}$ 考察的所有属性将其分到掌握组或未掌握组；在此基础上估计项目参数 $\hat{s}_{jc}$ 和 $\hat{g}_{jc}$，并计算项目区分度 $\delta_{jc}=1-\hat{s}_{jc}-\hat{g}_{jc}$；最后通过搜索算法，将最大的 δ_{jc} 所对应的 $\boldsymbol{q}_{jc}$ 作为项目 j 的 q 向量。

该方法考虑项目区分两个极端被试组的能力，容易推广到其它 CDMs，但 CDMs 不同，被试分组不同，计算过程也有差异。总体上，δ 方法简单易行，但往往不能反映全体被试的信息。于是，de la Torre 和 Chiu (2016)提出 ς^2 方法，以反映项目区分所有被试的能力。

2.1.2　广义区分度指标：ς^2 方法

de la Torre 和 Chiu (2016)提出计算全体被试正确作答项目概率 $P_j(\boldsymbol{\alpha}_l)$ 的方差，称为广义区分度指标，以反映项目区分所有被试的能力，即 $\varsigma_j^2=\sum_{l=1}^{2^K}w(\boldsymbol{\alpha}_l)[P_j(\boldsymbol{\alpha}_l)-\bar{P}_j]^2$ 。其中， $w(\boldsymbol{\alpha}_l)$ 表示被试 $\boldsymbol{\alpha}_l$ 的后验概率， $P_j(\boldsymbol{\alpha}_l)$ 表示被试 $\boldsymbol{\alpha}_l$ 的正确作答概率， $\bar{P}_j$ 表示所有被试平均的正确作答概率。针对项目 j ， ς^2 方法首先计算 $\boldsymbol{q}_{jc}$ 下正确作答概率的方差 ς_{jc}^2 ，然后选择方差占比(即 $\varsigma_{jc}^2/\max\{\varsigma_{j1}^2, \varsigma_{j2}^2,\ldots,\varsigma_{j,2^K-1}^2\}$)最大且考察属性最少的 $\boldsymbol{q}_{jc}$ 作为它的 q 向量。与 δ 方法相比，ς^2 方法利用了更多信息且具有一般性，总体上对 Q 矩阵的修正结果也更好(Wang et al., 2018)。有研究指出，ς^2 方法易受样本量影响，若样本量过小会大大降低它的表现(汪大勋　等, 2019)。

2.1.3　最优项目区分度方法的简评

δ 和 ς^2 方法都假设正确 q 向量应该使项目的区分度最优，并分别从项目区分极端被试组和所有被试组的能力建构项目区分度指标。虽然它们分别基于 DINA 和 G-DINA 模型提出，但都很容易推广至其它 CDMs。

δ 和 ς^2 方法对所有项目进行修正，项目越多，耗时越长。鉴于此，涂冬波等人(2012)提出先筛选再修正的 γ 方法。γ 方法首先根据项目参数的值筛选出可能存在冗余(缺失)属性的项目，进一步根据掌握组和未掌握组被试在每个属性上掌握概率之差判定该属性是否冗余(缺失)，以此为据修正项目 q 向量。DINA 模型下，γ 方法的 KS 判准率比 δ 方法更高，二者对 Q 矩阵的修正效果相当(涂冬波　等, 2012)。然而，γ 方法可能漏掉参数合理但 q 向量有误的情况。

除了先筛选后修正的方法外，研究者还运用不同搜索算法提高搜索效率。常用的搜索算法有穷举搜索算法、顺序搜索算法和逐步搜索算法。特别地，Ma 和 de la Torre (2020)在多级评分项目

的序列 G-DINA 模型中首先基于 ς^2 方法确定第一个必要属性对应的 q 向量，然后在逐步搜索算法中根据 Wald 统计量检验修订前后两个 q 向量的模型拟合是否存在显著差异，进而确定项目 q 向量，称为 stepwise 法。其中，Wald 统计量为 $W = [R \times P_j]'[R \times V_j \times R']^{-1}[R \times P_j]$，$P_j$ 表示不同属性模式的被试在项目 j 上的正确作答概率矩阵，V_j 为信息矩阵的逆矩阵 $I(P_j)^{-1}$，R 为限制性矩阵，具体算法见 Ma 和 de la Torre (2020)。基于信息矩阵的 Wald 统计量在认知诊断模型拟合、项目水平的模型拟合和项目功能差异检验中都有广泛的研究与应用(刘彦楼 等, 2019; 刘彦楼 等, 2016; Liu, Xin, et al., 2019; Liu, Andersson, et al., 2019; Liu, Yin, et al., 2019; Liu et al., 2016)。

此外，汪大勋等人(2019)和 Wang 等人(2020)将两个 q 向量的反应似然比 $D = -2\ln\left(\frac{L(\boldsymbol{Y} \mid \hat{\boldsymbol{\beta}}, \boldsymbol{Q}_{j_{q_{before}}})}{L(\boldsymbol{Y} \mid \hat{\boldsymbol{\beta}}, \boldsymbol{Q}_{j_{q_{after}}})}\right)$ 进行 χ^2 检验以确定更拟合的 q 向量。研究表明，似然比检验方法最优，Stepwise 法次之，ς^2 和 δ 法最差(Ma & de la Torre, 2020; Wang et al., 2018; Wang et al., 2020)。

δ 和 ς^2 方法基于绝对最优项目区分度指标确定 q 向量；γ 方法基于属性区分度进行效应量检验；stepwise 法和似然比检验重点探讨搜索算法与差异检验量在 Q 矩阵估计(修正)中的表现。事实上，考察其它区分度指标(如优势比、认知诊断区分度和属性区分度指标)、探索其它反映 q 向量合理性的统计检验指标，以及探讨如何提高搜索算法的准确性和速度都是值得研究的问题。

2.2 最优观察反应分布与期望反应分布的拟合：绝对拟合指标方法

这类方法的关键是建构反映观察反应概率和期望反应概率分布的差异性或一致性指标。S 统计量(Liu et al., 2012)、似然比 D^2 统计量(喻晓锋, 罗照盛, 高椿雷 等, 2015)和残差方法(chen, 2017)是这类方法的代表。

2.2.1 S 统计量方法

该方法的核心在于构建期望(观察)正确作答概率分布矩阵 $T_{u \times 2^K}$ ($\rho_{u \times 1}$)。其中，u 表示单个项目和不同项目组合的个数，T 矩阵的元素表示不同属性掌握模式的被试正确作答某个项目或某些项目组的概率，ρ 矩阵的元素表示实际正确作答项目或项目组的人数比例。令 $P_{2^K \times 1}$ 代表知识状态的先验分布，则理论上有 $T * P = \rho$。对项目 j，S 统计量方法选择使 $T_{jc} \cdot P$ 与 ρ 的欧氏距离最小的 $\boldsymbol{q}_{jc}$ 作为它的 q 向量。

注意到，T 矩阵的行数随项目个数的增加而极速增加，计算量也随之增大。于是，Liu 等人(2012)建议 T 矩阵至少包含 1 阶、2 阶到 K+1 阶不同项目组合。该方法考虑到项目与项目组合的反应，利用信息多，要求样本量大，计算量也随着项目和属性个数的增加而增大。

S 统计量方法提出之初，备受欢迎。例如，Xiang (2013)将 Q 矩阵元素视为连续变量，运用 S 统计量方法进行估计。该方法通过定义连续变量与属性的关系模型，并与阈值(0.5)相比转化为二值计分 Q 矩阵，但表现不佳。又如，喻晓锋、罗照盛、秦春影等人(2015)在 KS、$\boldsymbol{\beta}$ 与 Q 矩阵的联合估计中运用 S 统计量方法估计 Q 矩阵，还创新性地考察了当测验属性个数界定错误时该方法的表现。再如，杭丹丹(2020)将 S 统计量方法推广到多级计分项目，发现当 Q 矩阵失误率较小(5%)和被试足够多(N=4000)时该方法才具有较高的修正率。

2.2.2 似然比 D^2 统计量方法

喻晓锋、罗照盛、高椿雷等人(2015)将似然比 G^2 统计量(McKinley & Mills, 1985)应用于 CDMs，称为似然比 D^2 统计量。它表达了观察反应分布和期望反应分布的一致性，即

$$D_j^2 = 2\sum_{l=1}^{2^K}\left[\left(r_{lj}\log_{10}\frac{f_{lj}}{(1-s_{jc})^{\eta_{ljc}} g_{jc}^{1-\eta_{ljc}}} + (N_l - r_{lj})\log_{10}\frac{1-f_{lj}}{s_{jc}^{\eta_{ljc}}(1-g_{jc})^{1-\eta_{ljc}}}\right)\right] \quad (1)$$

其中，N_l、r_{lj} 和 f_{lj} 分别表示第 l 组中被试总数、观察正确作答项目 j 的人数和比例。s_{jc} 和 g_{jc} 分别表示 DINA 模型中 $\boldsymbol{q}_{jc}$ 下的项目参数。于是，该方法选择使 D^2 最小的 $\boldsymbol{q}_{jc}$ 作为项目 j 的 q 向量。他们基于部分 q 向量已知的项目(称为基础题)循环修正其余项目的 q 向量，直到前后两次的 Q 矩阵相同或达到最大迭代次数为止。结果表明，该算法与 S 统计量方法相比更省时、修正率更高。但是当样本量和基础题数量较少时，该算法可能不收敛，同时显著降低 Q 矩阵修正率。

2.2.3　残差方法

Chen (2017)根据期望反应分布和观察反应分布中项目 j 与项目 j' 的作答反应和正确作答人数，分别建立了基于相关的残差(记为 $r_{jj'}$)和基于对数比的残差(记为 $l_{jj'}$)，即

$$r_{jj'} = |Z[Cor(Y_j, Y_{j'})] - Z[Cor(\tilde{Y}_j, \tilde{Y}_{j'})]| \quad (2)$$

$$l_{jj'} = \left|\log_{10}\left(\frac{N_{11}\cdot N_{00}}{N_{01}\cdot N_{10}}\right) - \log_{10}\left(\frac{\tilde{N}_{11}\cdot \tilde{N}_{00}}{\tilde{N}_{01}\cdot \tilde{N}_{10}}\right)\right| \quad (3)$$

其中，$N_{tt'}$ 和 $\tilde{N}_{tt'}$ 表示在项目 j 和 j' 上分别得 t 和 t' 分的实际人数和期望人数，$Z[\cdot]$ 表示皮尔逊相关系数的 Fisher 转换值。

残差方法包括四个步骤。首先对所有项目对的 Z 分数($Zr_{jj'}$ 或 $Zl_{jj'}$)的最大值进行显著性检验，初步判定测验 Q 矩阵是否存在错误。当测验水平 Q 矩阵有误时，若测验水平的均方根(Sr 或 Sl)超过临界值，则应考虑测验属性个数是否冗余或缺失，并在测验水平进行修正。反之，在项目水平进行修正，并将项目水平中最大的几个均方根(记为 $Sr_j(Sl_j)$)对应的项目作为待修正项目集。最后，通过调整 q 向量前后 $Sr_j(Sl_j)$ 的变化或 $Sr_j(Sl_j)$ 最大值的变化确定待修正项目的 q 向量。结果表明，基于 r 和 l 的两种策略都能有效检测 Q 矩阵的错误(Chen et al., 2013)并修正项目 q 向量，但在短测验中的效果较差(Chen, 2017)。残差方法提出的层层筛选、判断和修正思路，不仅考虑到属性层面和项目水平的可能错误，还可以在一定程度上提高修正效率。

2.2.4　最优观察反应分布和期望反应分布的拟合：相对拟合指标方法

−2LL、AIC 和 BIC 是常用的模型数据拟合指标。对于 $-2LL = -2\ln(\prod_{i=1}^{N}\sum_{l=1}^{2^K} L(\mathrm{Y}_i \mid \hat{\boldsymbol{\beta}}, \boldsymbol{\alpha}_l) w(\boldsymbol{\alpha}_l))$ 中的 $w(\boldsymbol{\alpha}_l)$ 而言，有研究采用先验概率(Chen et al., 2013)，也有研究采用后验概率(汪大勋 等, 2020)来计算。AIC 和 BIC 在−2LL 的基础上分别加上模型参数个数 m 的惩罚因子 $2m$ 和代表模型参数与被试人数的惩罚因子 $m\ln(N)$。这些方法依据候选 $\boldsymbol{q}_{jc}$ 估计的项目参数计算拟合指标，并选择具有最优拟合的属性模式作为项目的 q 向量。Chen 等人(2013)和汪大勋等人(2020)分别在不同实验条件下考察了它们在 Q 矩阵修正中的表现。结果表明，无论模型是否为真，BIC 在不同 q 向量错误类型中的表现都优于 AIC 和−2LL 方法(Chen et al., 2013)；对复杂多级评分模型，BIC 表现同样优于 AIC 和−2LL 方法(汪大勋 等, 2020)。

AIC 和 BIC 对复杂模型的参数个数和样本量进行了惩罚，Chen 等人(2015)则在 $\ln(\prod_{i=1}^{N}\sum_{l=1}^{2^K} L(\mathrm{Y}_i \mid \hat{\boldsymbol{\beta}}, \boldsymbol{\alpha}_l) w(\boldsymbol{\alpha}_l))$ 基础上分别加上项目参数的 L_1 (Lasso)惩罚函数和 SCAD (Smoothly Clipped Absolute Deviation)惩罚函数，称为正则化 Q 矩阵估计方法。他们指出与 L_1 惩罚函数相比，SCAD 惩罚函数的结果更优，且在小样本条件下仍有较高的估计准确率。Xu 和 Shang (2018)研究表明，先运用正则化 L_1 惩罚函数方法估计 Q 矩阵，然后根据 BIC 指标修正 q 向量，能提高 Q 矩阵估计准确率。

2.2.5　基于数据拟合方法的简评

S 统计量反映了正确作答项目与项目对的观察概率分布和期望概率分布的欧氏距离，似然比 D^2 统计量则是对所有被试组的不同作答反应的观察概率分布与期望概率分布之比的对数加权平均之和。类似地，Kang 等人(2019)和杨亚坤等人(2020)考察了近似误差均方根 $RMSEA_j = \sqrt{\sum_{t=0}^{1}\sum_{l=1}^{2^K} w(\boldsymbol{\alpha}_l)\left(P_j(\boldsymbol{\alpha}_l) - \frac{n_{jlt}}{N_{jl}}\right)^2}$ 的表现，发现 RMSEA 方法能有效地估计 Q 矩阵，短测验中的修正效果也优于 δ 法和非参数欧氏距离法。Yu 和 Cheng (2020)还提出加权的残差指标 $R = \sum_{i=1}^{N}\log_{10}\left(\frac{Y_{ij}-\eta_{ij}}{P(Y_{ij}|\boldsymbol{\alpha}_i)}\right)^2$，研究表明该方法比 S 统计量方法更简单，修正结果更好。

S 统计量从项目总体上判断所有被试组的观察分布与期望分布的差异；似然比 D^2 统计量、RMSEA 和加权残差 R 方法从被试总体判断项目 j 的观察分布与期望分布的差异；而残差方法则是基于项目对的相关或对数比的绝对值误差。虽然似然比 D^2 统计量基于 DINA 模型提出，仍可以推广到其它 CDMs。总体上，绝对拟合指标的方法不受 CDMs 的限制，能较好地估计(修正)Q 矩阵。

与绝对拟合指标方法相比，相对拟合指标方法计算更简单，在实践中表现较好，应用广泛(汪大勋 等, 2019)。特别地，只有当项目参数个数随 $\boldsymbol{q}_{jc}$ 不同而不同时，AIC、BIC 与−2LL 才不同，否则它们是等价的。今后可以探索建立认知诊断模

型数据拟合指标或者考察其它拟合统计量(如 M_2 和 l_z 统计量)在 Q 矩阵估计(修正)中的表现。另外，正则化方法根据模型拟合程度和模型复杂性探索潜在维数，适用于测验考察的属性个数未知的情况，为估计 Q 矩阵提出了新的情境与思路。

2.3 基于参数估计的方法

将 Q 矩阵元素视为待估参数进行估计仍是一种有效的 Q 矩阵估计(修正)方法。目前研究主要考察了极大似然估计(maximum likelihood estimation, MLE)、边际极大似然估计(marginal maximum likelihood estimation, MMLE)方法(Wang et al., 2018)和贝叶斯估计(Chung, 2019; Chen et al., 2018; DeCarlo, 2012; Templin & Henson, 2006)在 Q 矩阵估计(修正)方面的表现。

2.3.1 MLE 和 MMLE 方法

MLE 和 MMLE 都采用 EM 算法循环估计 KS、$\boldsymbol{\beta}$ 和 Q 矩阵，直到收敛。具体地，在第 h 次迭代中，首先基于 $\hat{Q}^{(h-1)}$ 和作答反应矩阵估计 $\hat{\boldsymbol{\beta}}^{(h)}$ 和 $\hat{\boldsymbol{\alpha}}^{(h)}$。在此基础上, MLE 和 MMLE 方法分别计算似然 $L(\boldsymbol{Y}_j \mid \hat{\boldsymbol{\alpha}}_i^{(h)}, \hat{\boldsymbol{\beta}}_j^{(h)}, \boldsymbol{q}_{jc}) = \prod_{i=1}^{N} P_{jc}(\hat{\boldsymbol{\alpha}}_i^{(h)})^{Y_{ij}} (1 - P_{jc}(\hat{\boldsymbol{\alpha}}_i^{(h)}))^{1-Y_{ij}}$ 和边际似然 $ML(\boldsymbol{Y}_j \mid \hat{\boldsymbol{\beta}}_j^{(h)}, \boldsymbol{q}_{jc}) = \prod_{i=1}^{N} \sum_{l=1}^{2^K} L(Y_{ij} \mid \boldsymbol{\alpha}_l, \boldsymbol{\beta}_j^{(h)}, \boldsymbol{q}_{jc}) w(\boldsymbol{\alpha}_l \mid \boldsymbol{Y}_i, \boldsymbol{\beta}_j^{(h)}, \hat{\boldsymbol{q}}_{jc}^{(h-1)})$。接下来分别选择使 L 或 ML 取最大值的属性模式作为项目 j 的 q 向量。两者的区别在于是否利用了 KS 的后验分布信息。结果表明, MLE 倾向于保留正确的 q 向量。总体上, MMLE 方法优于 MLE。它们均优于 δ、γ 和 ς^2 方法(Wang et al., 2018)。

2.3.2 贝叶斯方法

Q 矩阵估计(修正)中常用的贝叶斯方法有期望后验估计(expected a posteriori, EAP)和马尔科夫链蒙特卡洛(Markov Chain Monte Carlo, MCMC)方法。首先, Templin 和 Henson (2006)以及 DeCarlo (2012)在不同条件下运用了 EAP 方法修正 Q 矩阵元素。他们令 Q 矩阵中不确定元素 q_{jk} 取值为 1 的概率为 $P(q_{jk}=1)$，并假设它服从 $Beta(a,b)$ 的先验分布。当从参数为 $P(q_{jk}=1)$ 的伯努利分布中随机抽样得到 $\tilde{q}_{jk}$ 时，有 $P(q_{jk}=1|\tilde{q}_{jk}) \sim Beta(a+\tilde{q}_{jk}, b+1-\tilde{q}_{jk})$，于是，后验期望为 $E(P(q_{jk}=1) \mid \tilde{q}_{jk}) = \frac{a+\tilde{q}_{jk}}{a+b+1}$。对不同 $\tilde{q}_{jk}$ 的后验期望求加权平均值后再取整确定 q_{jk} 的值。结果表明，当 Q 矩阵中除不确定元素外其余元素都正确时，EAP 方法能准确估计所有不确定元素(Templin & Henson, 2006)，但当其余元素有错时会显著降低 EAP 的修正率(DeCarlo, 2012)。

其次，MCMC 是基于后验分布进行抽样获取参数估计值的方法。Chen 等人(2018)和 Chung (2019)的研究表明，MCMC 方法能有效估计 Q 矩阵，但易受样本量和属性间相关的影响。特别地，MCMC 方法中运用 Metropolis Hastings (MH)或约束性吉布斯(Constrained Gibbs, CGibbs)抽样比运用吉布斯(Gibbs)抽样对 Q 矩阵估计更准确，且 CGibbs 抽样在小样本条件下的结果也比较好(Chen et al., 2018)。此外它们均优于 Chen 等人(2015)的正则化惩罚方法。

2.3.3 基于参数估计方法的简评

MLE 和 MMLE 方法都是常用的参数估计方法，简单易懂，但多次使用 EM 算法往往比较耗时。贝叶斯参数估计方法基于先验分布获取待估参数的后验分布，然后用后验分布的均值或样本均值作为估计值。Templin 和 Henson (2006)以及 DeCarlo (2012)提出的 EAP 方法耗时短，但需要预先指定 Q 矩阵的不确定元素，限制了其使用范围。Chen 等人(2018)和 Chung (2019)提出的 MCMC 方法在多种实验条件下的表现较好，但随着 K 的增加，需要更长的链才能收敛。总体上，基于参数估计的方法是一类重要的 Q 矩阵估计(修正)方法。事实上，这类方法常常需要对项目参数、KS 与 Q 矩阵进行联合估计，而它们的估计精度又相互影响。于是，如何表征它们的估计误差、如何在估计过程中结合估计误差都是参数估计方法中有价值的研究问题。

3 非参数 Q 矩阵估计(修正)方法

非参数方法不依赖 CDMs，也不运用参数化方法分析项目和被试特征。与参数化方法相比，非参数 Q 矩阵估计(修正)方法可利用的信息更少。非参数情境下，研究者们主要基于统计分析视角，通过最小观察反应向量和理想反应向量的距离、分析异常作答反应或视为因素结构进行因素分析研究 Q 矩阵估计(修正)问题。

3.1 最小观察反应向量与理想反应向量距离的方法

这类方法首先运用非参数认知诊断方法分析

被试的 KS, 并据此计算被试 i 在项目 j 候选 $\boldsymbol{q}_{jc}$ 下的理想反应 η_{ijc}，进而获得所有被试观察反应 $\boldsymbol{Y}_j$ 与理想反应 η_{jc} 的距离 d，最后选择使 d 最小的属性模式作为项目 j 的 q 向量。d 可采用欧氏距离 $\sum_{i=1}^{N}(\boldsymbol{Y}_{ij}-\eta_{ijc})^2$ (Barnes, 2010; Chiu, 2013; 杭丹丹, 2020)、海明距离 $\sum_{i=1}^{N}\mathrm{I}(\boldsymbol{Y}_{ij}\neq\eta_{ijc})$ (汪大勋, 高旭亮, 韩雨婷 等, 2018)或曼哈顿距离 $\sum_{i=1}^{N}\left|\boldsymbol{Y}_{ij}-\eta_{ijc}\right|$ (刘芯伶, 2020)来计算。不难发现, 它们在二级评分项目下是等价的。其中, Barnes (2010)将 q 向量作为连续变量, 以 0.1 为间隔变化 Q 矩阵元素的值, 当欧氏距离降至预设标准时获得 Q 矩阵, 但该方法判准率不高; Chiu (2013)和汪大勋、高旭亮、韩雨婷等人(2018)、杭丹丹(2020)和刘芯伶(2020)分别在二级和多级评分项目中考察这些距离的表现。结果表明, 样本容量、KS 的估计精度和基础题的数量都是影响这类方法的重要因素。

除将 KS 与 $\boldsymbol{q}_{jc}$ 对比获得无猜测无失误的理想反应外, 汪文义等人(2018)提出应基于被试观察反应获得该被试的理想反应。具体地, 该方法首先确定 $\boldsymbol{q}_{jc}$ 由可达矩阵 R 中哪些列通过布尔“或”运算而来, 然后根据被试在 R 阵中这些列所对应的项目上的观察反应进行布尔“与”运算获得其在 $\boldsymbol{q}_{jc}$ 上的理想反应, 记为 η_{jc}。结果表明, 在非参数欧氏距离判别法中运用这种方法获取理想反应时, 当可达阵 R 的项目参数小于 0.2, 待标项目的参数小于 0.3 时, Q 矩阵的元素返真率达 0.9 以上。运用该方法获取理想反应需要分析属性层级结构并以 R 阵为基础, 遗憾的是, 研究者并未比较 η_{ijc} 与 η_{jc} 两种理想反应的优劣。

3.2　最小异常反应指标方法

借鉴属性层级一致性指标(hierarchy consistency index, HCI) (Cui, 2007)的思想, 汪大勋、高旭亮、蔡艳等人(2018)构建了项目一致性指标(item consistency criterion, ICC), 用于表示对 $\boldsymbol{q}_{jc}$ 而言, 具有父级、子级和同级关系的项目对间作答反应的一致性。ICC 计算公式如下：

$$ICC=1-\frac{2\sum_{i=1}^{N}(\sum_{g\in S_{gc}}x_{ij}(1-x_{ig})+\sum_{f\in S_{fc}}x_{if}(1-x_{ij})+}{M_{jc}}$$
$$\frac{\sum_{h\in S_{hc}}(x_{ij}(1-x_{ih})+x_{ih}(1-x_{ij})))}{M_{jc}} \quad (4)$$

根据属性向量的包含关系, 可得项目 j 在 $\boldsymbol{q}_{jc}$ 下的子级、父级和同级项目集合 S_{gc}、S_{fc} 和 S_{hc}，并令 x_g、x_f 和 x_h 表示对应集合的项目反应。ICC 考虑了三种异常反应模式：项目 j 上答对但在子级项目上答错, 即 $x_j(1-x_g)$；项目 j 上答错但在父级项目上答对; 项目 j 上答对(错)但在同级项目上答错(对)。M_{jc} 表示比较的总次数。

实验表明, 当基础题大于 8 个时, ICC 方法的估计成功率接近 100%。总体上, 它在所有实验条件下均能较好地估计 Q 矩阵, 但不容易区分考察 K 个和 $K-1$ 个属性的项目(汪大勋, 高旭亮, 蔡艳等, 2018)。随后, 刘芯伶(2020)将 ICC 方法推广至多级计分项目, 发现其对 Q 矩阵的修正效果不及多级计分的曼哈顿距离方法和 stepwise 方法。此外, Wang 等人(2018)假设 $\boldsymbol{\alpha}_l$ 类被试的答对比例高于答错比例, 则 $\boldsymbol{q}_j$ 应属于 $\boldsymbol{\alpha}_l$ 的子集, 反之, $\boldsymbol{q}_j$ 属于 $\boldsymbol{\alpha}_l$ 的补集, 提出了通过集合交运算和差运算来修正 Q 矩阵的交叉方法(intersection and difference, ID)。结果表明, 当 N 较大时, ID 方法能较好地识别和修正错误的 q 向量, 优于 δ 法、欧氏距离法和 MLE 方法。ICC 和 ID 方法都突破了非参数距离的整体分析思想, 试图结合项目的反应过程、项目与被试的交互分析 q 向量, 为 Q 矩阵估计(修正)提供了新的思路和方法。

3.3　因素分析方法

Close (2012)和汪文义等人(2015)都提出应用因素分析方法探索项目与属性的结构。前者对主成分分析法获得的成分间相关系数矩阵进行分析获得 Q 矩阵; 后者对项目对的四分相关矩阵进行探索性因素分析得到初始 Q 矩阵, 然后运用 MLE 或 ID 方法对初始 Q 矩阵进行修正。结果表明, 基于四分相关矩阵的方法在样本量较小和参数较大的情况下均能有效估计 Q 矩阵。鉴于猜测和失误会带来反应误差进而影响四分相关矩阵, 汪文义等人(2020)通过极端高低分组估计 s、g, 并结合观察反应计算各类期望反应的人数, 改进四分相关矩阵的计算, 进而提高了 Q 矩阵估计的准确性。

3.4　非参数方法的简评

上述三类非参数方法本质上也是分布拟合和参数估计方法。首先, 基于 CDMs 的参数化方法分析实际作答反应概率分布和期望反应概率分布间的拟合情况; 基于统计分析的非参数方法通过求离散的观察反应向量与理想反应向量间的距离或异常反应指标来表达观察反应和理想反应间的

拟合程度。其次, 参数化方法中将 Q 矩阵视作未知模型参数进行估计; 而非参数方法则基于大量反应数据, 将 Q 矩阵元素视作项目与潜在属性之间的因子结构进行因素分析, 实质上是对因子结构的估计。再次, 大部分非参数方法主要适用于二级评分项目, 也没有与其它方法进行比较(如 Chiu, 2013; 汪大勋, 高旭亮, 韩雨婷 等, 2018; 汪文义 等, 2018)。因此, 今后可以考察其它距离判别法(如兰氏距离或杰卡德距离)、也可以开发非参项目特征指标、建构分布拟合特征进一步探索更多适用于短测验和小样本的非参数 Q 矩阵估计(修正)方法。

4 研究展望

Q 矩阵反映了项目特征, 其正确性决定着认知诊断结果的准确性。它是认知诊断微观评估的基础, 在认知诊断中具有举足轻重的作用。本文通过梳理国内外相关研究, 探讨 Q 矩阵估计(修正)方法的分类, 从而明晰相关研究的发展脉络。在此基础上, 详细介绍了各种方法的思路和步骤, 并分析了它们的特点、联系和区别。总体上, 当前研究一方面基于 CDMs, 从最优项目区分度、模型数据拟合和参数估计视角提出许多参数化 Q 矩阵估计(修正)方法; 另一方面基于统计分析, 依据最小观察反应向量和理想反应向量距离、最小异常反应指标和因素分析提出多种非参数 Q 矩阵估计(修正)方法。然而, 当前研究还缺乏对已有方法进行系统深入的比较、复杂测验情景的研究与应用, 未来还有待针对多种不同测验条件多角度开发 Q 矩阵估计(修正)方法。

4.1 系统比较 Q 矩阵估计(修正)方法

目前, 大部分研究基于模拟实验对相同类别的方法进行比较, 较少对不同类别的方法进行交叉比较, 而且多集中在 δ 、ς^2 、非参数欧氏距离、RMSEA、MLE 和 MMLE 方法间的比较(Kang et al., 2019; Wang et al., 2018)。因此, 今后有必要对不同类方法进行系统比较; 另外, 复杂测验条件的研究不多。今后应全面考察项目质量、被试特征、测验条件对各方法的影响, 以期为 Q 矩阵估计(修正)方法的实践应用提供方法、技术支持与实验证据。此外, 当前研究以模拟实验为主, 还应在实际反应数据中考察这些方法的表现。

表 1 Q 矩阵估计(修正)方法分类

分类标准		特点	方法	实验目的
参数化方法	最优项目特征	项目区分度	δ 法、ς^2 法、stepwise 法	修正
		属性区分度	γ 法	修正
	最优模型数据拟合	绝对拟合指标	S 统计量、多级计分的 S 统计量方法	修正
			非线性惩罚估计法	估计
			RMSEA 法、加权残差 R 法、残差方法	修正
			似然 D2 统计量方法	估计
		相对拟合指标	−2LL、AIC、BIC 方法	修正
			正则化极大似然估计方法	估计
			TS 法、LR-S 法、LR-E 法	估计
	参数估计	基于 EM 算法	MLE 和 MMLE 方法	修正
		基于贝叶斯的方法	EAP 方法	修正
			MCMC 方法	估计
非参数方法	最小观察反应与理想反应的距离	依据欧氏距离、海明距离、曼哈顿距离	欧氏距离法、多级计分的欧氏距离法	修正
			海明距离方法	修正
			曼哈顿距离方法	估计
	最小异常反应指标	依据被试在父级、子级和同级项目上的反应	ICC 法	估计
			多级计分的 ICC 法	修正
		依据 KS 与项目反应关系	ID 法	修正
	因素分析	将 Q 矩阵元素视作项目与潜在属性间的因子结构	主成分分析法	估计
			四分相关矩阵法	估计

4.2　多角度研究 Q 矩阵估计(修正)方法

无论通过最优项目特征、模型数据拟合还是参数估计获取 Q 矩阵, 都假定项目参数、KS 等是准确的。事实上, 它们与 Q 矩阵估计(修正)是紧密相连的统一体, 彼此相互影响估计精度。因此, 如何引入项目参数与 KS 的估计误差以提高 Q 矩阵估计精度具有重要意义。另外, 大部分研究以 DINA 模型、融合模型或 G-DINA 模型为基础, 在一定程度上限制了方法的使用条件。于是, 今后应基于更一般的认知诊断模型、反应时模型和高阶认知诊断模型等复杂模型探索已有方法的特点或提出新方法。最后, 开发项目特征指标、建构数据拟合指标、利用反应过程信息、结合多种思路和方法或提出新的研究视角都是深入研究 Q 矩阵估计(修正)方法可行的思路。

4.3　探讨不同条件下的 Q 矩阵估计(修正)方法

随着考试和评价方式的多样化, 测验形式越来越丰富, 测验条件越来越复杂。未来研究应关注多级评分项目、混合测验模型、属性多级、属性个数未知甚至 Q 矩阵元素为连续变量时 Q 矩阵的估计(修正)方法。探索如何将 Q 矩阵估计(修正)方法运用到在线标定中, 探讨联合标定 Q 矩阵和项目参数也是今后研究的重要方向。

参考文献

杭丹丹. (2020). *多级计分认知诊断评估中的 Q 矩阵验证方法与应用研究* (硕士学位论文). 江西师范大学, 南昌.

刘芯伶. (2020). *多级评分情境下 Q 矩阵修正的非参数方法* (硕士学位论文). 浙江师范大学, 金华.

刘彦楼, 辛涛, 李令青, 田伟, 刘笑笑. (2016). 改进的认知诊断模型项目功能差异检验方法——基于观察信息矩阵的 Wald 统计量. *心理学报, 48*(5), 588–598.

刘彦楼, 张倩萌, 郑宗军, 尹昊. (2019). 认知诊断模型中项目水平模型比较统计量的健壮性. *心理科学, 42*, 1251–1259.

涂冬波, 蔡艳, 戴海琦. (2012). 基于 DINA 模型的 Q 矩阵修正方法. *心理学报, 44*(4), 558–568.

汪大勋, 高旭亮, 蔡艳, 涂冬波. (2018). 一种非参数化的 Q 矩阵估计方法: ICC-IR 方法开发. *心理科学, 41*(2), 466–474.

汪大勋, 高旭亮, 蔡艳, 涂冬波. (2019). 一种广义的认知诊断 Q 矩阵修正新方法. *心理科学, 42*(4), 988–996.

汪大勋, 高旭亮, 蔡艳, 涂冬波. (2020). 基于类别水平的多级计分认知诊断 Q 矩阵修正: 相对拟合统计量视角. *心理学报, 52*(1), 93–106.

汪大勋, 高旭亮, 韩雨婷, 涂冬波. (2018). 一种简单有效的 Q 矩阵估计方法开发: 基于非参数化方法视角. *心理科学, 41*(1), 180–188.

汪文义, 高朋, 宋丽红, 汪腾. (2020). 带噪音预处理的改进探索性 Q 矩阵标定方法. *江西师范大学学报(自然科学版), 44*(2), 136–141.

汪文义, 宋丽红, 丁树良. (2015). 基于探索性因素分析的 Q 矩阵标定方法. *江西师范大学学报(自然科学版), 39*(2), 138–144+170.

汪文义, 宋丽红, 丁树良. (2018). 基于可达阵的一种 Q 矩阵标定方法. *心理科学, 41*(4), 968–975.

喻晓锋, 罗照盛, 高椿雷, 李喻骏, 王睿, 王钰彤. (2015). 使用似然比 D~2 统计量的题目属性定义方法. *心理学报, 47*(3), 417–426.

喻晓锋, 罗照盛, 秦春影, 高椿雷, 李喻骏. (2015). 基于作答数据的模型参数和 Q 矩阵联合估计. *心理学报, 47*(2), 273–282.

杨亚坤, 朱仕浩, 刘芯伶. (2020). 基于项目拟合统计量 RMSEA 的 Q 矩阵估计方法. *心理技术与应用, 8*(1), 51–59.

Barnes, T. (2010). Novel derivation and application of skill matrices: The Q-matrix method. In C. Ramero, S. Vemtora, M. Pechemizkiy, & R. S. J. de Baker (Eds.), *Handbook of educational data mining* (pp. 159–172). Boca Raton, FL: Chapman & Hall.

Chen, J. S. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement, 41*(4), 277–293.

Chen, J. S., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140.

Chen, Y. H., Culpepper, S. A., Chen, Y. G., & Douglas, J. (2018). Bayesian estimation of the DINA Q-matrix. *Psychometrika, 83*(1), 89–108.

Chen, Y. Statistical analysis of Q-matrix based diagnostic classificatio X., Liu, J. C., Xu, G. J., & Ying, Z. L. (2015). n models. *Journal of the American Statistical Association, 110*(510), 850–866.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598–618.

Chung, M. T. (2019). A Gibbs sampling algorithm that estimates the Q-matrix for the DINA model. *Journal of Mathematical Psychology, 93*, 102275.

Close, C. N. (2012). *An exploratory technique for finding the Q-matrix for the DINA model in cognitive diagnostic assessment: Combining theory with data* (Unpublished doctorial dissertation). University of Minnesota.

Cui, Y. (2007). *The hierarchy consistency index: A person -fit statistic for the attribute hierarchy method* (Unpublished doctorial dissertation). University of Alberta, Edmonton.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*(6), 447–468.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253–273.

Kang, C. H., Yang, Y. K., & Zeng, P. H. (2019). Q-matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement, 43*(7), 527–542.

Liu, J. C., Xu, G. J., & Ying, Z. L. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*(7), 548–564.

Liu, Y. L., Andersson, B., Xin, T., Zhang, H. Y., & Wang, L. L. (2019). Improved Wald statistics for item-level model comparison in diagnostic classification models. *Applied Psychological Measurement, 43*, 402–414.

Liu, Y. L., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*, 3–26.

Liu, Y. L., Xin, T., Andersson, B., & Tian, W. (2019). Information matrix estimation procedures for cognitive diagnostic models. *British Journal of Mathematical and Statistical Psychology, 72*, 18–37.

Liu, Y. L., Yin, H., Xin, T., Shao, L. C., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology, 10*, 1137.

Ma, W. C., & de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential generalized DINA model. *The British Journal of Mathematical and Statistical Psychology, 73*(1), 142–163.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*(1), 49–57.

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68(1)*, 78–96.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Wang, D. X., Cai, Y., & Tu, D. B. (2020). Q-matrix estimation methods for cognitive diagnosis models: Based on partial known Q-matrix. *Multivariate Behavioral Research ()*.

Wang, W. Y., Song, L. H., Ding, S. L., Meng, Y. R., Cao, C. X., & Jie, Y. J. (2018). An EM-based method for Q-matrix validation. *Applied Psychological Measurement, 42*(6), 446–459.

Xiang, R. (2013). *Nonlinear penalized estimation of true Q-matrix in cognitive diagnostic models (*Unpublished doctorial dissertation). Columbia University, New York.

Xu, G. J., & Shang, Z. R. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association, 113*(523), 1284–1295.

Yu, X. F., & Cheng, Y. (2020). Data-driven Q-matrix validation using a residual-based statistic in cognitive diagnostic assessment. *British Journal of Mathematical and Statistical Psychology. 73,* 145–179.

Q-matrix estimation (validation) methods for cognitive diagnosis

LI Jia, MAO Xiuzhen, ZHANG Xueqin

(*Institute of Educational Sichuan Normal University, Chengdu* 610066, *China*)

Abstract: The *Q*-matrix, which represents important item characteristics by mapping attributes to items has been proved to be the core factor affecting the accuracy of cognitive diagnostic classification. It is of great value to study the methods of *Q*-matrix estimation (validation). First, the existing methods of *Q*-matrix estimation and validation are classified into 1) parameterized methods in the CDM perspective, including item differentiation, model-data fit index and parameter estimation; and 2) non-parametric methods in the statistical perspective, including the distance between observed and expected response vector, abnormal responses index and factor analysis. Then, these methods are introduced in terms of differences and relations, characteristics and performance. The advantages and disadvantages of each method are commented. At last, several future research directions are proposed. It is necessary to compare the *Q*-matrix estimation (validation) methods systematically under complex test conditions. It is also of vital importance to propose *Q*-matrix estimation (validation) methods by combining multiple thoughts and ways based on the calibration of knowledge state and parameter estimation error. It is meaningful to further study the *Q*-matrix estimation (validation) methods for polytomous scoring items, mixed test models, polytomous scoring attributes, unknown number of attributes and even continuous *Q*-matrix.

Key words: cognitive diagnosis models, *Q*-matrix, *Q*-matrix estimation (validation) methods, model-data fit, parameter estimation